2 Bayes' Rule

This topic is on Bayes' Rule and Bayesian Reasoning. Bayes' Rule is the key component in how to build likelihood functions, which when optimized are key to evaluating models based on data. Bayesian Reasoning is a much broader area that can go well beyond just finding the single "most optimal" model. This line of work, which this chapter will only introduce, reasons about the many possible models and can make predictions with this uncertainty taken into account.

2.1 Bayes' Rule

Given two events M and D Bayes' Rule states

$$\mathbf{Pr}(M \mid D) = \frac{\mathbf{Pr}(D \mid M) \cdot \mathbf{Pr}(M)}{\mathbf{Pr}(D)}.$$

This assumes nothing about the independence of M and D (otherwise its pretty uninteresting). To derive this we use

$$\mathbf{Pr}(M \cap D) = \mathbf{Pr}(M \mid D)\mathbf{Pr}(D)$$

and also

$$\mathbf{Pr}(M \cap D) = \mathbf{Pr}(D \cap M) = \mathbf{Pr}(D \mid M)\mathbf{Pr}(M),$$

combined to get $\mathbf{Pr}(M \mid D)\mathbf{Pr}(D) = \mathbf{Pr}(D \mid M)\mathbf{Pr}(M)$, from which we can solve for $\mathbf{Pr}(M \mid D)$. So Bayes' Rule is uncontroversially true; any frequentist vs. Bayesian debate is about how to model data and perform analysis, not the specifics of this rule.

Example: Checking Bayes' Rule

Consider two events M and D with the following joint probability table:

	$M = 1$	$M = 0$
$D = 1$	0.25	0.5
$D = 0$	0.2	0.05

We can observe that indeed $\mathbf{Pr}(M \mid D) = \mathbf{Pr}(M \cap D)/\mathbf{Pr}(D) = \frac{0.25}{0.75} = \frac{1}{3}$, which is equal to

$$\frac{\mathbf{Pr}(D \mid M)\mathbf{Pr}(M)}{\mathbf{Pr}(D)} = \frac{\frac{.25}{.2+.25}(.2 + .25)}{.25 + .5} = \frac{.25}{.75} = \frac{1}{3}.$$

But Bayes' rule is not very interesting in the above example. In that example, it is actually *more* complicated to calculate the right side of Bayes' rule than it is the left side.

Example: Cracked Windshield

Consider you bought a new car and its windshield was cracked, the event W. If the car was assembled at one of three factories A, B or C, you would like to know which factory was the most likely point of origin.

Assume that in Utah 50% of cars are from factory A (that is $\mathbf{Pr}(A) = 0.5$) and 30% are from factory B ($\mathbf{Pr}(B) = 0.3$), and 20% are from factory C ($\mathbf{Pr}(C) = 0.2$).

Then you look up statistics online, and find the following rates of cracked windshields for each factory – apparently this is a problem! In factory A, only 1% are cracked, in factory B 10% are cracked, and in factory C 2% are cracked. That is $\mathbf{Pr}(W \mid A) = 0.01$, $\mathbf{Pr}(W \mid B) = 0.1$ and $\mathbf{Pr}(W \mid C) = 0.02$.

We can now calculate the probability the car came from each factory:

- $\mathbf{Pr}(A \mid W) = \mathbf{Pr}(W \mid A) \cdot \mathbf{Pr}(A)/\mathbf{Pr}(W) = 0.01 \cdot 0.5/\mathbf{Pr}(W) = 0.005/\mathbf{Pr}(W)$.
- $\mathbf{Pr}(B \mid W) = \mathbf{Pr}(W \mid B) \cdot \mathbf{Pr}(B)/\mathbf{Pr}(W) = 0.1 \cdot 0.3/\mathbf{Pr}(W) = 0.03/\mathbf{Pr}(W)$.
- $\mathbf{Pr}(C \mid W) = \mathbf{Pr}(W \mid C) \cdot \mathbf{Pr}(C)/\mathbf{Pr}(W) = 0.02 \cdot 0.2/\mathbf{Pr}(W) = 0.004/\mathbf{Pr}(W)$.

We did not calculate $\mathbf{Pr}(W)$, but it must be the same for all factory events, so to find the highest probability factory we can ignore it. The probability $\mathbf{Pr}(B \mid W) = 0.03/\mathbf{Pr}(W)$ is the largest, and B is the most likely factory.

2.1.1 Model Given Data

In data analysis, M represents a 'model' and D as 'data.' Then $\mathbf{Pr}(M \mid D)$ is interpreted as the probability of model M given that we have observed D. A *maximum a posteriori* (or MAP) estimate is the model $M \in \Omega_M$ that maximizes $\mathbf{Pr}(M \mid D)$. That is

$$M^* = \arg\max_{M \in \Omega_M} \mathbf{Pr}(M \mid D) = \arg\max_{M \in \Omega_M} \frac{\mathbf{Pr}(D \mid M)\mathbf{Pr}(M)}{\mathbf{Pr}(D)} = \arg\max_{M \in \Omega_M} \mathbf{Pr}(D \mid M)\mathbf{Pr}(M).$$

Thus, by using Bayes' Rule, we can maximize $\mathbf{Pr}(M \mid D)$ using $\mathbf{Pr}(M)$ and $\mathbf{Pr}(M \mid D)$. We do not need $\mathbf{Pr}(D)$ since our data is given to us and fixed for all models.

In some settings we may also ignore $\mathbf{Pr}(M)$, as we may assume all possible models are equally likely. This is not always the case, and we'll come back to this. Thus we just need to calculate $\mathbf{Pr}(D \mid M)$. Then, in this setting $L(M) = \mathbf{Pr}(D \mid M)$ is called the *likelihood* of model M.

So what is a 'model' and what is 'data?' A *model* is usually a simple pattern which we think data is generated from, but then observed with some noise. Examples:

- The model M is a single point in $\mathbb{R}^d$; the data is a set of points in $\mathbb{R}^d$ near M.
- **linear regression:** The model M is a line in $\mathbb{R}^2$; the data is a set of points such that for each x-coordinate, the y-coordinate is the value of the line at that x-coordinate with some added noise in the y-value.
- **clustering:** The model M is a small set of points in $\mathbb{R}^d$; the data is a large set of points in $\mathbb{R}^d$, where each point is near one of the points in M.
- **PCA:** The model M is a k-dimensional subspace in $\mathbb{R}^d$ (for $k \ll d$); the data is a set of points in $\mathbb{R}^d$, where each point is near M.
- **linear classification:** The model M is a halfspace in $\mathbb{R}^d$; the data is a set of labeled points (with labels $+$ or $-$), so the $+$ points are mostly in M, and the $-$ points are mainly not in M.

Example: Gaussian MLE

Let the data D be a set of points in $\mathbb{R}^1 : \{1, 3, 12, 5, 9\}$. Let Ω_M be $\mathbb{R}$ so that the model is a point $M \in \mathbb{R}$. If we assume that each data point is observed with independent Gaussian noise (with $\sigma = 2$, so its pdf is described as $g(x) = \frac{1}{\sqrt{8\pi}} \exp(-\frac{1}{8}(M-x)^2)$. Then

$$\mathbf{Pr}(D \mid M) = \prod_{x \in D} g(x) = \prod_{x \in D} \left(\frac{1}{\sqrt{8\pi}} \exp(-\frac{1}{8}(M-x)^2) \right).$$

Recall that we can take the product $\Pi_{x \in D} g(x)$ since we assume independence of $x \in D$! To find $M^* = \arg\max_M \mathbf{Pr}(D \mid M)$ is equivalent to $\arg\max_M \ln(\mathbf{Pr}(D \mid M))$, the *log-likelihood* which is

$$\ln(\mathbf{Pr}(D \mid M)) = \ln \left(\prod_{x \in D} \left(\frac{1}{\sqrt{8\pi}} \exp(-\frac{1}{8}(M-x)^2) \right) \right) = \sum_{x \in D} \left(-\frac{1}{8}(M-x)^2 \right) + |D| \ln(\frac{1}{\sqrt{8\pi}}).$$

We can ignore the last term since it is independent of M. The first term is maximized when $\sum_{x \in D}(M-x)^2$ is minimized, which occurs precisely as $\mathbf{E}[D] = \frac{1}{|D|} \sum_{x \in D} x$, the mean of the data set D. That is, the maximum likelihood model is exactly the mean of the data D, and is quite easy to calculate.

2.2 Bayesian Inference

Bayesian inference focuses on a simplified version of Bayes's Rule:

$$\mathbf{Pr}(M \mid D) \propto \mathbf{Pr}(D \mid M) \cdot \mathbf{Pr}(M).$$

The symbol $\propto$ means *proportional to*; that is there is a fixed (but possibly unknown) constant factor c multiplied on the right (in this case $c = 1/\mathbf{Pr}(D)$) to make them equal: $\mathbf{Pr}(M \mid D) = c \cdot \mathbf{Pr}(D \mid M) \cdot \mathbf{Pr}(M)$.

However, we may want to use continuous random variables, so then strictly using probability **Pr** at a single point is not always correct. So we can replace each of these with pdfs

$$p(M \mid D) \propto f(D \mid M) \cdot \pi(M).$$

Each of these terms have common names. As above, the conditional probability or pdf $\mathbf{Pr}(D \mid M) \propto f(D \mid M)$ is called the *likelihood*. The probability or pdf of the model $\mathbf{Pr}(M) \propto \pi(M)$ is called the *prior*. And the left hand side $\mathbf{Pr}(M \mid D) \propto p(M \mid D)$ is called the *posterior*.

Again it is common to be in a situation where, given a fixed model M, it is possible to calculate the likelihood $f(D \mid M)$. And again, the goal is to be able to compute $p(M \mid D)$, as this allows us to evaluate potential models M, given the data we have seen D.

The main difference is a careful analysis of $\pi(M)$, the prior – which is not necessarily assumed uniform or "flat". The prior allows us to encode our assumptions.

Example: Average Height

Lets estimate the height H of a typical U of U student. We can construct a data set $D = \{x_1, \ldots, x_n\}$ by measuring the height of everyone in this class in inches. There may be error in the measurement, and we are an incomplete set, so we don't entirely trust the data.
So we introduce a prior $\pi(M)$. Consider we read that the average height of an full grown person is $\mu_M = 66$ inches, with a standard deviation of $\sigma = 6$ inches. So we assume

$$\pi(M) = N(66, 6) = \frac{1}{\sqrt{\pi 72}} \exp(-(\mu_M - 66)^2/(2 \cdot 6^2)),$$

is normally distributed around 66 inches.
Now, given this knowledge we adjust the MLE example from last subsection using this prior.

- *What if our MLE estimate without the prior (e.g. $\frac{1}{|D|}\sum_{x \in D} x$) provides a value of 5.5?* That means the data is very far from the prior. Usually this means something is wrong. We could find $\arg\max_M p(M \mid D)$ using this information, but that may give us an estimate of say 20 (that does not seem correct). A more likely explanation is a mistake somewhere: probably we measured in feet instead of inches!

Another vestige of Bayesian inference is that we not only can calculate the maximum likelihood model M^*, but we can also provide a posterior value for any model! This value is not an absolute probability (its not normalized, and regardless it may be of measure 0), but it is powerful in other ways:

- We can say (under our model assumptions, which are now clearly stated) that one model M_1 is twice as likely as another M_2, if $p(M_1 \mid D)/p(M_2 \mid D) = 2$.

- We can define a range of parameter values (with more work and under our model assumptions) that likely contains the true model.

- We can now use more than one model for prediction of a value. Given a new data point x' we may want to map it onto our model as $M(x')$, or assign it a score of fit. Instead of doing this for just one "best" model M^*, we can take a weighted average of all models, weighted by their posterior; this is "marginalization."

Weight for Prior. So how important is the prior? In the average height example, it will turn out to be worth only $(1/9)$th of one student's measurement. But we can give it more weight.

Example: Weighted Prior for Height

Lets continue the example about the height of an average U of U student, and assume (as in the MLE estimator example) the data is generated independently from a model M with Gaussian noise with $\sigma = 2$. Thus the likelihood of the model, given the data is

$$f(D \mid M) = \prod_{x \in D} g(x) = \prod_{x \in D} \left(\frac{1}{\sqrt{8\pi}} \exp(-\frac{1}{8}(\mu_M - x)^2) \right).$$

Now using that the prior of the model is $\pi(M) = \frac{1}{\sqrt{\pi 72}} \exp(-(\mu_M - 66)^2/72)$, the posterior is given by

$$p(M \mid D) \propto f(D \mid M) \cdot \frac{1}{\sqrt{\pi 72}} \exp(-(\mu_M - 66)^2/72).$$

It is again easier to work with the log-posterior which is monotonic with the posterior, using some unspecified constant C (which can be effectively ignored):

$$\begin{aligned} \ln(p(M \mid D)) &\propto \ln(f(D \mid M)) + \ln(\pi(M)) + C \\ &\propto \sum_{x \in D} \left(-\frac{1}{8}(\mu_M - x)^2) \right) - \frac{1}{72}(\mu_M - 66)^2 + C \\ &\propto -\sum_{x \in D} 9(\mu_M - x)^2 + (\mu_M - 66)^2 + C \end{aligned}$$

So the maximum likelihood estimator occurs at the average of 66 along with 9 copies of the student data.

Why is student measurement data worth so much more?

We assume the standard deviation of the measurement error is 2, where as we assumed that the standard deviation of the full population was 6. In other words, our measurements had variance $2^2 = 4$, and the population had variance $6^2 = 36$ (technically, this is best to interpret as the variance when adapted to various subpopulations, e.g., U of U students): that is 9 times as much.

If instead we assumed that the standard deviation of our prior is 0.1, with variance 0.01, then this is 400 times smaller than our class measurement error variance. If we were to redo the above calculations with this smaller variance, we would find that this assumption weights the prior 400 times the effect of each student measurement in the MLE.

In fact, a much smaller variance on the prior is probably more realistic since national estimates on height are probably drawn from a very large sample. And its important to keep in mind that we are estimating the *average height* of a population, not the *height of a single person* randomly drawn from the population. In the next topic (T3) we will see how averages of random variables have much smaller variance – are much more concentrated – than individual random variables.

So what happens with more data?

Lets say, this class gets really popular, and next year $1{,}000$ or $10{,}000$ students sign up! Then again the student data is overall worth more than the prior data. So with any prior, if we get enough data, it no longer becomes important. But with a small amount of data, it can have a large influence on our model.